DOCUMENT RESUME

ED 112 412 CS 202 266

AUTHOR Marzano, Robert J.
TITLE On the Validity of Analytic Ratings.
PUB DATE 75
NOTE 7p.; Unpublished study prepared at the Univ. of Colorado at Denver

EDRS PRICE MF-$0.76 HC-$1.58 Plus Postage
DESCRIPTORS *Composition (Literary); Educational Research; English Instruction; *Evaluation Methods; Higher Education; *Test Reliability; *Test Validity; Writing Skills; Written Language

ABSTRACT

The purpose of this study was to examine the reliability of the analytical method of grading essays in relation to the holistic method. It was hypothesized that the use of the analytic method to rate college composition papers produces high rater reliability at the expense of biasing the raters and thus lowering the validity of the grades. Six essays, all on the same topic, were used for the study. It was concluded that the analytic method of rating produces a higher reliability among raters than does the holistic method, but that the analytic method produces a lower validity for the grades on the papers than does the holistic method. On the basis of the study, the hypothesis that the analytic method lowers rater validity by introducing rater bias was logically, but not statistically, accepted. (RB)

Robert J. Marzano
Assistant Professor of Education
University of Colorado
Denver, Colorado

ON THE VALIDITY OF ANALYTIC RATINGS

The unreliability of quality grades for essays was dramatically illustrated by Paul Diederich in 1961. Diederich examined the ratings of essays made by 53 readers from six different professional areas: English, Law, Natural Science, Social Science, Writers/Editors, Business. Ninety four percent of the papers received seven or more of the nine possible grades: the median correlation between the readers was .31. The highest reader reliability of .41 was registered by those readers from the field of English.

Each of the six groups of Diederich's readers used what can be termed a holistic method of rating. Readers were given no specific intructions for rating papers. They were asked to judge a paper's quality based on their "overall" impression of it.

It has been suggested that Diederich's findings accurately represent the unreliability of essay grading in general. However, as Ebel and Damrin (1960) point out, if trained raters follow clearly articulated criteria, reader reliability can be increased, especially if rater teams are used to evaluate essays. The use of clearly articulated criteria for rating has been termed the "analytic" method. That the analytic method can improve the reliability of essay grading has been demonstrated by Follman and Anderson (1967) and others.

There is, however, one very basic, yet unanswered, question concerning the analytic method of rating. That is: Does the analytic method produce ratings that are as valid or more valid than the holistic method? This question has been indirectly raised by Magnusson (1967, p. 124) who writes that..."it happens occasionally

that high reliability, for instance in the form of agreement among different judges giving subjective ratings... is taken to be a sign of the ratings' validity... Such an agreement is not a sufficient basis for concluding high validity: It can arise because the judges have the same bias in common, and the ratings perhaps, express something entirely different from what was intended..."

On the basis of Magnusson's comments, it was hypothesized that the use of the analytic method produces high reliability at the expense of biasing the raters and thus lowering the validity. Formally it was hypothesized that the quality ratings of papers judged using the holistic method would correlate higher with a criterion rating, than the quality ratings of those same papers judged using an analytic method.

Six essays, all on the same topic, were used for the study. The criterion rating was extablished using "authority." Three university professors, all of whom teach composition, independently rated the papers. The judges were asked to use the method of rating they had personally found most accurate and efficient over the years. All three judges used a method that can be considered a cross between the holistic and analytic approaches. All judges had predetermined categories that they "kept in mind" while rating. The type and specificity of categories differed from judge to judge; when asked to define the categories one professor was fairly explicit, but the other two were very general in their descriptions. None of the judges assigned numeric weights to the categories but, instead, used them only as a guide for their overall rating. Hence, the criterion ratings were established using a technique which had some aspects of the analytic

method (loosely defined, predetermined categories) and some aspects of the holistic method (non-numeric, subjective ranking of papers for overall quality).

The inter-rater reliability (Kendall's Coefficient of Concordance) for the authority raters was .85. This was interpreted as an indication that the six papers represented different and recognizable levels of quality. The mean rank of the three authority ratings was used as the criterion rank for each paper.[1]

Eight subjects were assigned randomly to two groups of raters, four raters per group. All subjects were college seniors studying to be secondary English teachers. The raters in Group 1 were instructed to use the following holistic method for rating the six papers:

Figure 1 goes here

The raters in Group II were instructed to use the following analytic scale:

Figure 2 goes here

The inter-rater reliability (Coefficient of Concordance) was calculated for each group. The holistic group (Group I) had a reliability of .59; the analytic group (Group II) had a reliability of .70.

---

[1]Baker, Hardyck and Petrinovich (1966) have shown that the use of ordinal scales in the calculation of means and other statistics does not significantly violate the underlying mathematical or measurement assumptions of statistical analysis.

The papers from each group were then ranked (based on the mean ranking for each paper) and rank order correlations (rho) calculated between the criterion ranking and the ranking for each group. The holistic method produced a correlation of .80 with the criterion ranking; the analytic method produced a correlation of .47 with the criterion ranking. The two correlations were considered to be the validity indices for the holistic and analytic methods, respectively.

It was concluded that the analytic method of rating produces a higher reliability than the holistic method (.70 vs .59), but the analytic method produces a lower validity than the holistic method (.47 vs .80). On the basis of the study, the hypothesis that the analytic method lowers validity by introducing rater bias was logically (not statistically)[2] accepted.

From an intuitive point of view this is quite logical. The differences between good and poor writing are numerous and perhaps too complexly interrelated to be measured by the present store of analytic scales that utilize discrete categories. Until the characteristics of good, average and poor writing have been defined, it is futile to list criteria from which writing quality should be judged.

[2]To this writer's knowledge, there is no known sampling distribution for chance differences between rho coefficients calculated on the same population. Hence no test of significance was performed.

Figure 1

"Everyman's Scale"

Please evaluate the six essays you have been given. Rate each essay independently. In other words, rate the first essay, then rate the second essay, etc. There is no particular grade that each essay should receive. You evaluate each essay according to your own judgment as to what constitutes writing ability. Use your own judgment about the writing ability as indicated by each essay. Don't use any system other than your own judgment. When you have judged each paper, sort them into a pile according to their quality. The first paper should be the best of the group; the last paper should be the worst of the group.

Figure 2

Diederich Rating Scale

	Low		Middle		High
Ideas	2	4	6	8	10
Organization	2	4	6	8	10
Wording	1	2	3	4	5
Flavor	1	2	3	4	5
Usage	1	2	3	4	5
Punctuation	1	2	3	4	5
Spelling	1	2	3	4	5
Handwriting	1	2	3	4	5

Sum_____

REFERENCES

Baker, B., Hardyck, C. and Petrinovich, L., "Weak measurements vs. strong statistics: An empirical critique of S.S. Stevens' proscriptions on statistics," Educational and psychological measurement, 1966, 26, 291-309.

Diederich, P. and others, Factors in judgments of writing ability (Research Bulletin RB-61-15, Princeton, N.J.: ETS, 1961)

Ebel, R. and Damrin, D. "Tests and examinations," Encyclopedia of educational research, ed. C. Harris (New York: The Macmillan Co., 1960), pp. 1502-1517

Follman, J. and Anderson, J. "An investigation of the reliability of five procedures for grading English themes," Research in the teaching of English, 1967, 1, No. 2, 190-200.

Magnusson, D. Test theory (U.S.: Addison-Wesley Co., 1966)